Lingua Franca, March 2001, ABC

Languages on the Internet

A linguist can't help but be impressed by the Internet. It's an extraordinarily diverse medium, holding a mirror up to so many sides of our linguistic nature. The World Wide Web, in particular, offers a home to virtually all the styles which have so far developed in the written language – newspapers, scientific reports, bulletins, novels, poems, prayers – you name it, you'll find a page on it. Indeed, it's introducing us to new styles of written expression which none of us have ever seen before – animated language, in particular. Words which appear and disappear, in varying colours. Sentences which slide onto the screen and off again. Letters which dance before your eyes. The Web is truly part of a new linguistic medium – more dynamic than traditional writing, and more permanent than traditional speech. It's often been said, the Internet is a revolution – yes, indeed, but it's also a linguistic revolution.

I can give you a couple of quick examples, so that you see my point. Take e-mails. You receive a message which contains, say, three different points in a single paragraph. You can, if you want, reply to each of these points by taking the paragraph, splitting it up into three parts, and then responding to each part separately, so that the message you send back then looks a bit like a play dialogue. Then, your sender can do the same thing to your responses, and when you get the message back, you see his replies to your replies. You can then send the lot on to someone else for further comments, and when it comes back, there are now three voices present on the screen. And so it can go on – replies within replies within replies. And all looking exactly the same, in the same screen typography. There's never been anything like this in the history of human letter-writing. That's one reason why I say the Internet is a linguistic revolution.

Or say you're participating in a real-time Internet discussion group – a chatroom. You see on your screen messages coming in from all over the world. If there are 30 people in the room, then you could be seeing 30 different messages, all making various contributions to the theme, but often clustering into half a dozen or more sub-conversations. It's like being in a cocktail party where there are other conversations going on all around you. In the party, of course, you can't pay attention to them. In a chatroom you can't avoid them. It has never been possible before, in the history of human communication, to ~~listen~~ attend and talking to 30 people at once. Now you can. It's a revolution, all right.

But there's another reason for the revolutionary status of the Net – and this one may surprise you. It's because the Internet offers a home to *all* languages – as soon as their communities have a functioning computer technology, I mean. Its increasingly multilingual character has been the most notable change since it started out – not very long ago - as a totally English medium. There's a story the former US vice-president Al Gore tells. He was reporting the remark of the 8-year-old son of Kyrgyzstan's President Akayev, who told his father that he had to learn English. When asked why, the child apparently replied: 'Because, daddy, the computer speaks English.'

For many, indeed, the language of the Internet 'is' English. There was a headline in *The New York Times* in 1996 which said simply: 'World, Wide, Web: 3 English Words'. The article, by Michael Specter, went on to say: 'if you want to take full advantage of the Internet there is only one real way to do it: learn English'. He did acknowledge the arrival of other languages: 'As the Web grows', he said, 'the number of people on it who speak French, say, or Russian will become more varied and that variety will be expressed on the Web. That is why it is a fundamentally democratic technology', he said, 'but it won't necessarily happen soon.'

Well, the evidence is growing that this conclusion was wrong. With the Internet's globalization, the presence of other languages has steadily risen. By the mid-1990s, a widely quoted figure was that about 80% of the Net was in English. People were acknowledging the first major study of language distribution on the Internet, carried out in 1997 by an organization called, believe it or not, Babel, a joint initiative of the Internet Society and Alis Technologies. This showed English well ahead, but with several other languages entering the ring – notably German, Japanese, French, and Spanish.

Since then, the estimates for English have been steadily falling. Some commentators are now predicting that before long the Web (and the Internet as a whole) will be predominantly *non*-English, as communications infrastructure develops in Europe, Asia, Africa, and South America. Listen to these results of a recent Global Reach survey. They estimated that people with Internet access in non-English-speaking countries increased between 1995 and 2000 from 7 million to 136 million. In 1998, there was another surprise: the number of newly created Web sites *not* in English passed the total for newly created sites that *were* in English. And at a conference on Search Engine Strategies in London last year, Alta Vista were predicting that by 2002 less than 50% of the Web would be in English. In certain parts of the world, the local language is already dominant. According to one Japanese Internet author Yoshi Mikami, 90% of Web pages in Japan are already in Japanese.

The Web is increasingly reflecting the distribution of language presence in the real world, and many sites provide the evidence. There are thousands of businesses now doing their best to present a multilingual identity. For instance, the Belgian daily newspaper, *Le Soir*, is represented by no less than six languages, French, Dutch, English, German, Italian, and Spanish. Then there are now hundreds of major sites collecting all kinds of data on the languages themselves. Call up the font archive at the University of Oregon, for example: you'll find 112 printing fonts in their archives for over 40 languages. They have a nice sense of humour, too – because you'll also find some data there on alien languages, such as Klingon, and folklore languages, such as Elvish, which Tolkien invented for *Lord of the Rings*

Spend an hour hunting for languages on the World Wide Web and you'll find hundreds. Last year I spent a few days tracking down as many examples as I could find, for my book *Language and the Internet*. I found one site, called World Language Resources, which lists products for 728 languages. I found an African resource list which covered several local languages; Yoruba, for example, was illustrated by some 5000 words, along with proverbs, naming patterns, and greetings. Another site dealt with no less than 87 European minority languages. Some of the sites were very small in content, of course, but nonetheless extensive in range: one gave the Lord's Prayer in nearly 500 languages.

Nobody has yet worked out just how many languages have obtained a modicum of presence on the Web. I found over 1000 quite quickly. It's not difficult to find evidence of a Net presence for all the more frequently used languages in the world, and for a large number of minority languages too. I'd guess that about a quarter of the world's languages – that's about 1500 - have some sort of cyber existence now.

It's important to point out that in all these examples I'm talking about language presence in a real sense. These aren't sites which only analyse or talk about languages, from the point of view of linguistics or some other academic subject. They're sites which allow us to see languages as they are. In many cases, the total Web presence, in terms of number of pages, is as I've said quite small. The crucial point is that the languages *are* out there, even if they're represented by only a sprinkling of sites. It's the ideal medium for minority languages, after all. Imagine, if you were one of the speakers or supporters of an endangered language – an aboriginal language, say, or a language like my own Welsh, or one of the other Celtic languages – well, you're keen to give the language some publicity, to draw its plight to the attention of the world. Previously, you'd have had a terrible time. Think of the difficulty of attracting a newspaper article on the subject, or the cost of a newspaper advertisement. It would be virtually impossible to get a radio or television programme devoted to it. But now, with Web pages waiting to be used, and e-mail there at the cost of a phone call, you can get your message out in next to no time, in your own language – with a translation as well, if you want - and in front of a global audience whose potential size makes traditional media audiences look minuscule by comparison.

On the other hand, I have to recognize that developing a significant cyber-presence for a language isn't easy. There's a sort of 'critical mass' of Internet penetration which has to build up in a country, before a language develops a vibrant cyber-life. It's not much use, really, to have just one or two sites in a local language on the Web. People wanting to use or find out about the language would soon get bored. The number of sites has to build up until, suddenly, everybody's using them and adding to them and talking about them. That's a magic moment, and only a few hundred languages have so far reached it. In the jargon of the Internet, there needs to be lots of good 'content' in the local languages out there, and until there is, people will stay using the languages that have managed to accumulate content – English, in particular.

So the future of a multilingual Internet isn't guaranteed. It will all depend on how quickly new sites can build up a local language momentum. There are also a number of practical difficulties. Until quite recently there were real problems in using the characters of the keyboard to cope with the alphabetical diversity of the world's languages. Because it was the English alphabet that was the standard, only a very few non-English accents and diacritics could be handled. If it was a foreign word with some strange-looking accent marks, the Internet software would simply ignore them, and assume they weren't important. This can still happen – but things have moved on a great deal since then. First, the basic set of keyboard characters, the so-called ASCII set, was extended, so that the commoner non-English diacritics could be included. But even then it only allowed up to 256 characters – and there are far more letter shapes in the world than that. Just think of the array of shapes you find in Arabic, Hindi, Chinese, Korean, and the many other languages which don't use the Latin alphabet. Today, a new coding system, the UNICODE system, is much more sophisticated: it allows the representation on screen of over 65,000 characters. That should be plenty – but the implementation of this system is still in its infancy.

My feeling is that the future looks good for Web multilingualism, and a number of influential people seem to share this view. Ned Thomas, for instance, is editor of a bulletin called *Contact* – it's the quarterly publication of the European Bureau of Lesser Used Languages. In an editorial last year he said: 'It is not the case ... that all languages will be marginalized on the Net by English. On the contrary, there will be a great demand for multilingual Web sites, for multilingual data retrieval, for machine translation, for voice recognition systems to be multilingual.' And Tyler Chambers, the creator of various Web language projects, agrees: 'the future of the Internet', he says, 'is even more multilingualism and cross-cultural exploration and understanding than we've already seen.' I agree. The Web offers a World Wide Welcome for global linguistic diversity.